

eirasagree

Paulo Sergio Panse Silveira¹ and Jose Oliveira Siqueira²

¹Department of Pathology, School of Medicine, University of Sao Paulo, São Paulo, SP, Brazil - email: silveira@usp.br, ORCID: 0000-0003-4110-1038

²Department of Pathology, School of Medicine, University of São Paulo, São Paulo, SP, Brazil - email: siqueira@usp.br, ORCID: 0000-0002-3357-8939

August 2, 2023

Contents

1	What is this package	1
2	Instructions for package installation	2
2.1	Dependences	2
2.2	Download	2
2.3	Installation	2
3	A first example	3
4	Demonstrations	5
4.1	Assumptions and conditions	5
4.2	Notation	5
4.3	Test of accuracy (location shift, systematic bias)	6
4.4	Test of precision (equality of variances)	10
4.5	Test of reliability ¹ (Deming regression)	16
5	Translations	27
6	Getting started with eirasagree	31

The package **eirasagree** implements all methods applied to:

Silveira PSP, Vieira JE, Siqueira JO. Is Bland-Altman plot method useful without inference for accuracy, precision, and agreement?,

accepted in June 20, 2023 to be published in the *Revista de Saúde Pública*. It is also available for download as preprint from Cornell University arXiv.org at <https://doi.org/10.48550/arXiv.2108.12937>.

This package is available for download from Harvard Dataverse at <https://doi.org/10.7910/DVN/AGJPZH>.

1 What is this package

The Bland and Altman plot method² is widely recognized and applied as a graphical approach to assess the comparability of different quantitative measurement techniques. Its primary objective is often to replace a traditional technique with a new one that is less invasive or less expensive. However, despite its widespread use, the Bland and Altman plot method is frequently misinterpreted due to the lack of appropriate inferential statistical support.

Common alternatives like Pearson's correlation or ordinal least-squares linear regression also fail to identify the weaknesses of each measurement technique. To address this limitation and provide inferential statistical support for assessing equivalence between measurement techniques, we propose three nested tests based on structural regressions. These tests evaluate the equivalence of structural means (accuracy), the equivalence of structural variances (precision), and the concordance with the structural bisector line (reliability of measurements from the same subject). We employ both analytical methods and robust bootstrapping approaches to obtain reliable inferential statistics.

In addition to the analytical methods, we have implemented graphical outputs that adhere to the principles established by Bland and Altman. This allows for effective visual communication while ensuring the incorporation of robust and appropriate inferential statistics to verify full equivalence between two measurement techniques. By decomposing equivalence into accuracy, precision, and reliability, our approach facilitates the identification of specific areas that require improvement in order to develop new measurement techniques.

2 Instructions for package installation

2.1 Dependences

This package depends on two other packages available in Harvard Dataverse, in addition to other packages available through CRAN:

- **eirasdata**, <https://doi.org/10.7910/DVN/DLQTPH>
- **eiras**, <https://doi.org/10.7910/DVN/TBBAVU>

which must be installed in this order.

Both packages are available for download in LEAP - Laboratory for Evaluation and Analysis in Psychometrics at <https://dataverse.harvard.edu/dataverse/LEAP>.

It also depends on some traditional R packages: `car`, `DescTools`, `estimatr`, `graphics`, `grDevices`, `MASS`, `textutils`, `rmarkdown`, and `knitr`.

2.2 Download

The package **eirasagree** is available for download from Harvard Dataverse at <https://doi.org/10.7910/DVN/AGJPZH>. This repository contains:

- **Install_and_Demonstrations.pdf** (this text) with installation instructions, package use examples, and mathematical demonstrations of the statistical tests.
- **eirasagree.1.0.6.tar.gz** is a compiled, ready-to-install file, with which all eirasagree package examples can be replicated.
- **sourcecodes.zip** for developers. It contains the R source codes and function documentation for package generation

2.3 Installation

For researchers who are interested in directly applying these procedures, installing the package archive file **eirasagree\1.0.6.tar.gz** will suffice, which includes all the necessary components, such as raw data, functions, and documentation with examples of use.

To install under RStudio:

- download the package file **tar.gz** in a local computer,
- access **Tools, Install Packages...**,
- change the option Install from: **Package Archive File (.tar.gz)**,
- browse to find the local package archive file (**tar.gz**),
- proceed to the installation.

To install using command line (e.g., Linux):

- download the package file **tar.gz** to a folder (e.g., `/home/your_user/Downloads`) in a local computer,
- open a terminal (e.g., depending on your system, a usual shortcut is **CTRL+ALT+T**),
- change to the folder where is the downloaded file (e.g., `cd /home/your_user/Downloads`),
- open R as superuser (e.g., `sudo R`),
- proceed to the installation (e.g., `install.packages("eirasagree_1.0.6.tar.gz")`).

3 A first example

The simplest way to analyze a dataframe is to use the function `AllStructuralTests`:

```
eirasagree::AllStructuralTests(eirasdata::PlasmaVolume,
                              reference.cols=2,
                              newmethod.cols=3,
                              alpha=0.05, out.format="html")
```

This single procedure generates the file `Nadler_x_Hurley.html`, named after the second and third columns names of a dataframe. In this case, the `eirasdata::PlasmaVolume` is a data frame available from `eirasdata` package (see section ‘Instructions for package installation’, page 2).

It can be replaced by any other dataframe. The numbers attributed to `reference.cols` and `newmethod.cols` correspond to the measurements of the reference and surrogate techniques, respectively.

More than one column can be indicated when a technique is applied more than once to the same subjects; in this case, this function will adjust for the corrected computation. For instance:

```
out <- eirasagree::AllStructuralTests(eirasdata::PEFR,
                                     reference.cols=c(1:2),
                                     newmethod.cols=c(3:4),
                                     alpha=0.05, out.format="html")
```

In both examples, an `analysis` subfolder is automatically created, and an HTML file is generated with a name based on the reference and new method columns. This HTML file serves as a comprehensive report, applying all the package functions in the appropriate order.

More examples are provided at the end of this text in the section titled “Getting started with `eirasagree`” (page 31). Additionally, there are other

examples embedded in the package documentation. After installing the package, you can give it a try:

```
> library(eirasagree)
> ?eirasagree::AllStructuralTests
```

4 Demonstrations

The proposed statistical method for the verification of equivalence between measurement techniques involves three statistical tests that are executed independently but conceptually nested: the equivalence of structural means (accuracy), the equivalence of structural variances (precision), and concordance with the structural bisector line (reliability in measurements obtained from the same subject) using analytical methods and a robust approach through bootstrapping. The third test is of greater importance, as the test of reliability between measurements obtained from the same individual using different measurement techniques only makes sense if both techniques are equal in terms of absence of bias (i.e., equal accuracy) and variance (i.e., equal precision). This verification of agreement with the bisector line should be performed using Deming regression, rather than conventional ordinary least squares regression, as described below.

4.1 Assumptions and conditions

- The outcome variable is interval-scaled.
- Using bootstrapping methods eliminates the need to assume normality and homoscedasticity.
- ONLY one population defined by inclusion and exclusion criteria.
- Measurement technique A is taken as *Reference or Gold Standard*.
- Measurement technique B is taken as *New, candidate method* (due to being less costly, less invasive, more efficient, or faster).
- The main tests are “acceptance” tests rather than hypothesis rejection tests: the sample size required to achieve a prospective power of 90% should be determined.
- Strict agreement requires linear relationship passing through the origin between methods A and B: bisector line of the first quadrant of the Cartesian plane.

4.2 Notation

The mathematical notation and the order of data for the reference and putative, candidate measurement techniques affect the writing of equations. Here, we will adopt the conventions applied at the NCSS manuals³:

- X : true/latent value of the measurement from the new method B
- Y : true/latent value of the measurement from the reference method A
- x : observed/manifest value of the measurement from the new method

- \bar{x} : arithmetic mean of observed values of x
- y : observed/manifest value of the measurement from the reference method
- \bar{y} : arithmetic mean of observed values of y
- ϵ_i and δ_i are the measurement errors with zero mean and are independent of each other, as well as independent of X and Y , respectively.
- $x_i = X_i + \epsilon_i$
- $y_i = Y_i + \delta_i$
- $i = 1, 2, \dots, n$ observational units
- The variances $V(\epsilon)$ and $V(\delta)$ are not necessarily equal.
- Ratio of variances: $\lambda = \frac{V(\delta)}{V(\epsilon)}$

For all subsequent tests, there are two null hypotheses (H_0). The structural H_0 corresponds to a conceptual hypothesis that is linked to the true, latent, structural values that are not directly observable. Therefore, it is necessary to establish a theorem that provides a computable, functional null hypothesis. These operational procedures involve modified ordinary least square linear regressions (OLS) that employ statistical techniques to establish a connection between structural values and procedures based on observable values.

These theorems were primarily developed within a theoretical framework and, due to mathematical complexities, remained relatively unknown and were not widely utilized by applied researchers. To the best of our knowledge, our implementation represents the first comprehensive integration of these scattered statistical theoretical results, which span from 1879 to 2015⁴⁻¹¹. Specifically, the theorem demonstrating accuracy was established by Hedberg and Ayres in 2015⁵, precision was addressed by Shukla in 1973⁶, and agreement with the true bisector line was explored by Kummell in 1879¹⁰; subsequently, these concepts became known as Deming regression and were revisited by Creasy in 1956⁴ and Watson and Petrie in 2021¹², among others.

The subsequent sections provide analytical descriptions of the accuracy, precision, and bisector agreement tests, along with illustrative R scripts. The main text also includes the versions of these tests that involve bootstrapping, which are implemented in the `erasagree` package.

4.3 Test of accuracy (location shift, systematic bias)

The measurement errors have an expected value (mean) of zero:

$$E[\epsilon] = E[\delta] = 0. \quad (1)$$

Therefore, the observed means are equal to the true means:

$$\begin{aligned} E[x] &= E[X + \delta] = E[X] + E[\delta] = E[X] \\ E[y] &= E[Y + \epsilon] = E[Y] + E[\epsilon] = E[Y] \end{aligned} \quad (2)$$

Hence, if both methods exhibit equal accuracy (with a difference equal to zero), then there is no bias, indicating that both techniques yield identical true means:

$$bias = E[x] - E[y] = E[X] - E[Y] = 0 \Rightarrow E[X] = E[Y]. \quad (3)$$

The null hypothesis is, therefore, the absence of bias:

$$\begin{aligned} \text{structural } H_0^1 : E[X] &= E[Y] \\ \text{functional } H_0^1 : \mu_d &= \mu_x - \mu_y = 0 \end{aligned} \quad (4)$$

Hedberg and Ayers⁵ demonstrated that the functional approach requires the use of a linear regression between $y - x$ (the individual measurement biases) and $x_i - \bar{x}$ (centering the measurements of method B). The intercept of this regression represents the mean bias between the candidate and reference measurement techniques. The proposed OLS regression between $y - x$ and $x - \bar{x}$, along with the corresponding t statistics to test $H_0^1 : \mu_d = 0$, are as follows:

$$\begin{aligned} y_i - x_i &= \mu_d + \beta(x_i - \bar{x}) + u_i, i = 1, 2, \dots, n \\ t &= \frac{\bar{y} - \bar{x}}{\sqrt{(1 - r_{xy}^2) \frac{s_y^2}{n}}} \sim t_{n-2} \end{aligned} \quad (5)$$

To make the procedures clearer, we illustrate with a numerical example. Suppose we have two measurement methods, x and y , applied to 15 individuals (Table 1).

Table 1: Hypothetical values of two measurement methods (x and y) applied to 15 individuals.

individual	x	y
1	5.3	6.2
2	8.9	9.4
3	4.4	5.7
4	9.4	11.0
5	1.9	3.1
6	6.6	7.2
7	2.4	4.5
8	3.9	5.2
9	7.7	8.3
10	4.8	6.5
11	6.7	7.8
12	0.2	2.6
13	8.4	9.9
14	5.6	6.7
15	8.2	9.8

Figure 1 illustrates the statistical maneuver of subtracting each value x from the mean value \bar{x} , resulting in a translation of the regression line. This maneuver does not affect the slope of the regression, but it impacts the intercept μ_d , representing the bias between the methods. When there is no rejection of $H_0^1: \mu_d = 0$, it indicates that both measurement techniques are equally accurate.

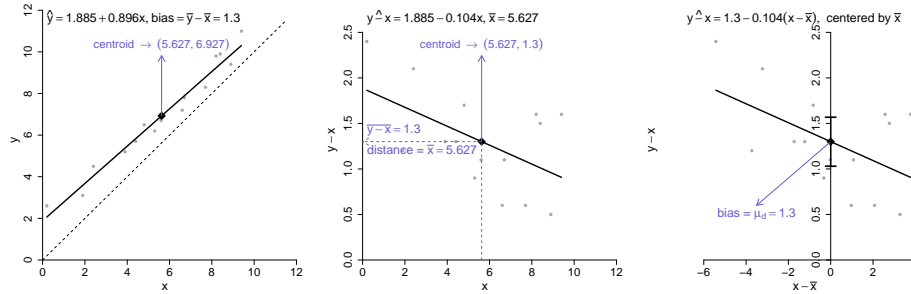


Figure 1: Illustration of Hedberg and Ayers' (2015) theorem using ordinary least square regressions from left to right: the first regression is the regular regression, the second replaces the y axis with $y - x$, and the third also replaces the x axis with $x - \bar{x}$. This clever maneuver effectively recovers the regression intercept, which serves as an indicator of the bias between the original values of x and y . The presence of bias can be statistically verified by testing $H_0: \mu_d = 0$.

The example depicted in Figure 1 also serves to emphasize that the regular OLS regression may not be a reliable estimator of equivalence between measurement techniques. Observe that the Pearson's correlation coefficient is high ($r = 0.983$) when computed by:

```
print(s.reg <-summary(estimatr::lm_robust(y ~ x))
r <- s.reg $r.squared^0.5
cat("\nPearson's r = ",r,sep="")
```

which results in:

```
Call:
estimatr::lm_robust(formula = y ~ x)

Standard error type: HC2

Coefficients:
              Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
(Intercept)  1.8848     0.33966   5.549 9.402e-05  1.1510  2.619 13
x             0.8961     0.05841  15.341 1.045e-09  0.7699  1.022 13

Multiple R-squared:  0.9659 ,      Adjusted R-squared:  0.9633
F-statistic: 235.4 on 1 and 13 DF,  p-value: 1.045e-09

Pearson's r = 0.9827929
```

However, there is statistical evidence of bias when testing $H_0: \mu_d = 0$,

computed with:

```
dv <- y - x
iv <- x - mean(x,na.rm=TRUE)
print(summary(estimatr::lm_robust(iv~dv)))
```

in which iv is the independent variable ($x - \bar{x}$) and dv is the depend variable ($y - x$), obtaining:

```
Call:
estimatr::lm_robust(formula = dv ~ iv)

Standard error type: HC2

Coefficients:
              Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
(Intercept)  1.3000     0.12586  10.329 1.234e-07  1.0281  1.57190 13
iv           -0.1039     0.05841  -1.779 9.857e-02 -0.2301  0.02226 13

Multiple R-squared:  0.2758 ,      Adjusted R-squared:  0.2201
F-statistic: 3.166 on 1 and 13 DF,  p-value: 0.09857
```

The statistics of interest is the coefficient line at the Intercept, showing a significant bias of 1.3 with a confidence interval of 95% [1.0281, 1.5719] (thus, not containing zero, as depicted in Figure 1, right panel) and $p = 1.234 \times 10^{-7}$.

The graphical representation of this test shows the 95% confidence band for the difference of population means, $E[y] - E[x] = 0$, obtained with this regression between $y - x$ and $x - \bar{x}$. The non-rejection of the null hypothesis corresponds to the point (0,0) located within the limits of this band. The confidence interval is represented in Figure 2 by a vertical line at the coordinate $x - \bar{x} = 0$. This figure can be obtained using the following script:

```
x.rac <- seq(min(iv), max(iv), length.out=1e3)
y.rac <- rac$coefficients[1,1] + rac$coefficients[2,1]*x.rac
sig.level <- 0.05
qF <- (2/n)*qf(1-sig.level/2, 2, n-2)
iv.mean <- mean(iv,na.rm=TRUE)
iv.sd <- sd(iv,na.rm=TRUE)
z2 <- ((x.rac - iv.mean)/iv.sd)^2
var.res <- rac$res_var
y.lwr <- y.rac - sqrt(qF*(1+z2)*var.res)
y.upr <- y.rac + sqrt(qF*(1+z2)*var.res)
y.rac0 <- rac$coefficients[1,1]
z2.0 <- ((-iv.mean)/iv.sd)^2
y.lwr.0 <- y.rac0 - sqrt(qF*(1+z2.0)*var.res)
y.upr.0 <- y.rac0 + sqrt(qF*(1+z2.0)*var.res)
minx <- min(c(0,x.rac),na.rm=TRUE)
maxx <- max(c(0,x.rac),na.rm=TRUE)
miny <- min(c(0,y.rac,y.lwr,y.upr),na.rm=TRUE)
maxy <- max(c(0,y.rac,y.lwr,y.upr),na.rm=TRUE)
sunflowerplot(iv,dv,
               xlim=c(minx,maxx),ylim=c(miny,maxy),
               main="Test of accuracy",
               xlab="x - mean(x)",
               ylab="y - x",
               rotate=TRUE,
               size=.1,
               col="gray",
               seg.col="gray",
               seg.lwd=.8,
               axes=FALSE)
```

```

axis(1)
axis(2)
lines(x.rac, y.rac, col="darkgray")
lines(x.rac, y.upr, lty=2)
lines(x.rac, y.lwr, lty=2)
# H0
abline(v=0, lty=3, lwd=0.9)
lines(rep(0,2), c(y.lwr.0, y.upr.0), lwd=2)
points(0,0,pch=19)

```

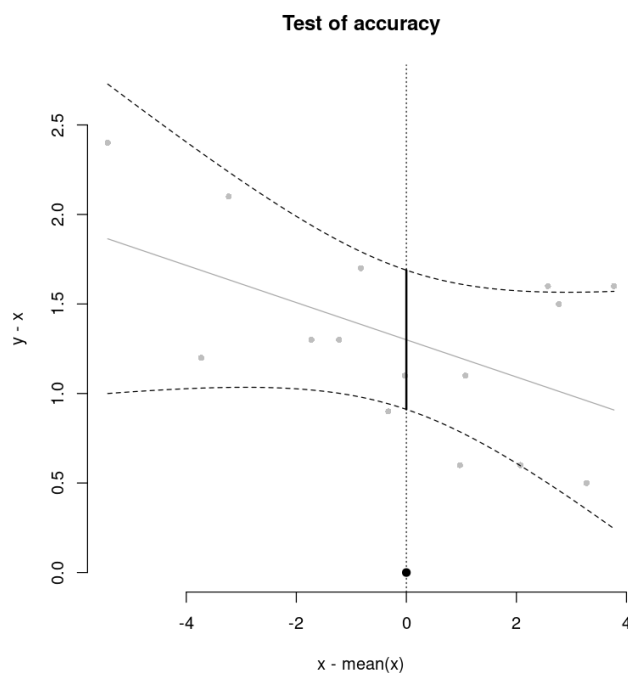


Figure 2: Graphical representation of the accuracy test with rejection of the null hypothesis. The bias is represented by the point (0,0) located outside the confidence interval, which is represented by a solid black line.

4.4 Test of precision (equality of variances)

If the latent values of measurements X and Y are equal (inherently equivalent methods), $X = Y = T$, then $V[T] = \sigma_T^2$ represents the variability of the common latent value T of the measurement, i.e., the inherent variance of what is being measured, disregarding measurement errors.

In addition, each measurement technique has a specific precision (noise variance):

- $V[\delta] = \sigma_\delta^2$: variability of the measurement error for method A

- $V[\epsilon] = \sigma_\epsilon^2$: variability of the measurement error for method B

The inherent variance of the measured quantity has to be significantly greater than the specific measurement precisions for both measurement instruments ($\sigma_T^2 \gg \sigma_\delta^2$ and $\sigma_T^2 \gg \sigma_\epsilon^2$). Otherwise, the measurement instruments are, by definition, not useful and no performance tests should be required.

Assuming that the measurement errors are independent of each other and the common latent value of the measurement, the following covariances (C) are null:

$$C[\delta, \epsilon] = C[T, \delta] = C[T, \epsilon] = 0 \quad (6)$$

It is a well-known fact that the covariance of the sum is always equal to the sum of the covariances. In other words, if random variables a , b , and c are dependent or independent with finite means and variances, then the covariance between the sum $a+b$ and c is equal to the sum of the covariances between a with b , and b with c : $C[a+b, c] = C[a, b] + C[b, c]$. Therefore, considering equation 6, it can be observed that the covariance between the observable measurements of the two methods represents the inherent variance of the measured quantity:

$$\begin{aligned} C[y, x] &= C[T + \delta, T + \epsilon] \\ C[y, x] &= C[T, T] + C[T, \delta] + C[T, \epsilon] + C[\delta, \epsilon] \\ C[y, x] &= C[T, T] \\ C[y, x] &= \sigma_T^2 \end{aligned} \quad (7)$$

Variance is also defined as the covariance of a variable with itself, thus:

$$\begin{aligned} V[x] &= C[x, x] \\ V[x] &= C[T + \epsilon, T + \epsilon] \\ V[x] &= C[T, T] + C[T, \epsilon] + C[T, \epsilon] + C[\epsilon, \epsilon] \\ V[x] &= C[T, T] + C[\epsilon, \epsilon] \\ V[x] &= \sigma_T^2 + \sigma_\epsilon^2 \end{aligned} \quad (8)$$

$$\begin{aligned} V[y] &= C[y, y] \\ V[y] &= C[T + \delta, T + \delta] \\ V[y] &= C[T, T] + C[T, \delta] + C[T, \delta] + C[\delta, \delta] \\ V[y] &= C[T, T] + C[\delta, \delta] \\ V[y] &= \sigma_T^2 + \sigma_\delta^2 \end{aligned} \quad (9)$$

We conclude that the total variance of each of the two observable measurements can be decomposed into two components: the inherent variance of the measured quantity and the specific measurement errors. In this context,

from equations 7, 8, and 9, the Pearson correlation between the two observable measurements is expressed by:

$$\rho_{xy} = \frac{C[y, x]}{\sqrt{V[x]V[y]}} = \frac{\sigma_T^2}{\sqrt{(\sigma_T^2 + \sigma_\delta^2)(\sigma_T^2 + \sigma_\epsilon^2)}} \quad (10)$$

If the precision of both methods is equal ($\sigma_\delta^2 = \sigma_\epsilon^2$), then the Pearson correlation coefficient ρ_{xy} can be expressed as $\frac{\sigma_T^2}{\sigma_T^2 + \sigma_\delta^2}$ or $\frac{\sigma_T^2}{\sigma_T^2 + \sigma_\epsilon^2}$, which is also known as the intraclass correlation coefficient (ICC) or reliability ratio.

In the case of an excellent ICC, it is expected that $\rho_{xy} \approx 1$, which occurs when the inherent variance of the measured quantity (σ_T^2) greatly exceeds the specific measurement errors (σ_δ^2 and σ_ϵ^2). This highlights the importance of the inherent variance (σ_T^2) dominating over the measurement errors (σ_δ^2 and σ_ϵ^2), which stems from our initial assumption. For instance, if the measurement instruments introduce errors, even equal errors, that are comparable to the variance of the measured quantity, then $\rho \approx 0.5$ at best, indicating unsatisfactory precision of the instruments. Consequently, in order to achieve satisfactory and equal precision for both measurement techniques, the following two conditions should be distinguished:

1. $\sigma_\delta^2 = \sigma_\epsilon^2 = \sigma_m^2$, $\sigma_T^2 \gg \sigma_m^2$
2. $\sigma_\delta^2 \neq \sigma_\epsilon^2$, although $\sigma_T^2 \gg \sigma_\delta^2$ and $\sigma_T^2 \gg \sigma_\epsilon^2$

The objective is to determine whether condition 1 led to a correlation close to 1, as this condition includes the requirement of equal specific precisions ($\sigma_\delta^2 = \sigma_\epsilon^2$). Note that if the specific precisions are equal, the total variances are also equal, i.e., $V[x] = \sigma_T^2 + \sigma_\epsilon^2 = \sigma_T^2 + \sigma_\delta^2 = V[y]$ (see equations 8 and 9).

Thus, the following null hypothesis can be proposed to perform the structural test of equality of precisions between the two methods:

$$\begin{aligned} \text{structural } H_0^2 : \lambda &= \frac{V[\delta]}{V[\epsilon]} = 1 \\ \text{functional } H_0^2 : \rho_{y+x, y-x} &= 0 \end{aligned} \quad (11)$$

Therefore, the corresponding linear regression between $y - x$ (the individual measurement biases) and $x + y$ (or $\frac{x + y}{2}$, which is a linear transformation that does not qualitatively change the regression), is:

$$y_i - x_i = \alpha + \beta(x_i + y_i) + \nu_i, i = 1, 2, \dots, n \quad (12)$$

Shukla in 1973⁶, who Shoukri later revised in 2010¹³, demonstrated that the slope of this regression represents the difference in precision between the candidate and reference measurement techniques.

It can be verified by the covariance between the sum and difference of the two observable measurements, which is:

$$\begin{aligned}
 C[y+x, y-x] &= C[y, y] + C[y, -x] + C[x, y] + C[x, -x] \\
 C[y+x, y-x] &= C[y, y] - C[y, x] + C[x, y] - C[x, x] \\
 C[y+x, y-x] &= C[y, y] - C[x, x] \\
 C[y+x, y-x] &= V[y] - V[x] \\
 C[y+x, y-x] &= \sigma_T^2 + \sigma_\delta^2 - \sigma_T^2 - \sigma_\epsilon^2 \\
 C[y+x, y-x] &= \sigma_\delta^2 - \sigma_\epsilon^2
 \end{aligned} \tag{13}$$

Consequently, the Pearson correlation between the sum and difference of the two observable measurements is:

$$\rho_{y+x, y-x} = \frac{C[y+x, y-x]}{\sqrt{V[y+x]V[y-x]}} = \frac{\sigma_\delta^2 - \sigma_\epsilon^2}{\sqrt{V[y+x]V[y-x]}} \tag{14}$$

If $\sigma_\delta^2 = \sigma_\epsilon^2$ (the condition of equal precisions), then $\rho_{y+x, y-x} = 0$, and vice-versa.

It is also known that the slope β of the linear regression presented in equation 12 is:

$$\beta = \rho_{y+x, y-x} \sqrt{\frac{V[y+x]}{V[y-x]}} \tag{15}$$

Thus, when $\rho_{y+x, y-x} = 0$, which represents the condition of equal precisions, $\beta = 0$. By the transitive property, if $\sigma_\delta^2 = \sigma_\epsilon^2$ (the condition of equal precisions), then $\beta = 0$, and vice versa. The statistical test can be used to test either the functional hypothesis $H_0^2 : \rho_{y+x, y-x} = 0$ (equation 11) or, equivalently, $H_0^2 : \beta = 0$.

Figure 3 returns to the numerical example of Figure 1 to test $H_0^2 : \beta = 0$.

The statistical test can be computed by:

```

meanxy <- (x+y)/2
diffyx <- y-x
print(rprc <-summary(estimatr::lm_robust(diffyx ~ meanxy)))

```

obtaining:

```

Call:
estimatr::lm_robust(formula = diffyx ~ meanxy)

Standard error type: HC2

Coefficients:
              Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
(Intercept)  1.88442    0.42658   4.418 0.0006949  0.9629  2.80598 13
meanxy      -0.09311    0.06561  -1.419 0.1794069 -0.2349  0.04864 13

Multiple R-squared:  0.2005 ,      Adjusted R-squared:  0.139
F-statistic: 2.014 on 1 and 13 DF,  p-value: 0.1794

```

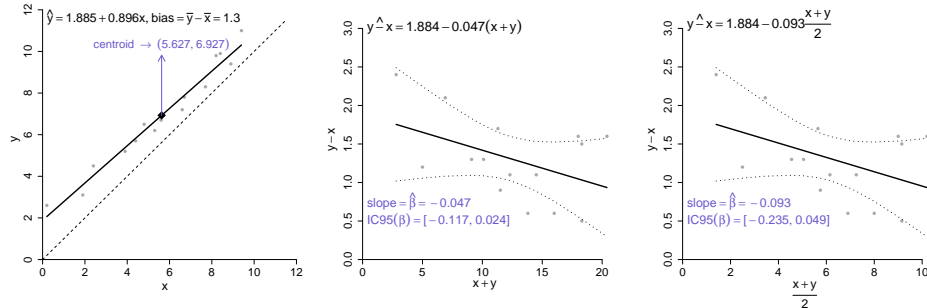


Figure 3: Illustration of Shukla/Shoukri^{6,13} theorems using ordinary least square regressions from left to right: the first regression is the regular regression, the second replaces the y axis with $y - x$ and x axis with $x + y$, and the third modifies the x axis to $\frac{x+y}{2}$, showing that it does not change the regression quality but has the advantage to preserve the order of magnitude of x and y original values. This maneuver effectively accesses the difference of precisions in such a way that null slope (β) corresponds to $\lambda = 1$ (i.e., the precision of both methods, $V[\delta]$ and $V[\epsilon]$, are comparable). Equality of precisions can be statistically verified by testing $H_0 : \beta = 0$.

The statistic of interest is the coefficient line at the meanxy (the line containing the slope estimates). In this example, there is no statistical evidence of a difference between precisions, as indicated by the $\beta = -0.09311$ coefficient with a confidence interval of $[-0.2349, 0.04864]$ (which includes zero) and a corresponding p-value of 0.1794. Graphically, as depicted in the central and right panels of Figure 3, a hypothetical horizontal line can be accommodated within the 95% confidence band (shown as dotted lines), indicating that it is possible for the population curve to be horizontal.

The graphical representation of the precision test shows the 95% confidence band obtained from a regression between $y - x$ and $\frac{x+y}{2}$. The non-rejection of the null hypothesis corresponds to a horizontal line segment located within the bounds of this band. Figure 4, similar to Figure 3 (right panel), can be generated using the following script:

```
x.rprc <- seq(min(meanxy), max(meanxy), length.out=1e3)
y.rprc <- rprc$coefficients[1,1] + rprc$coefficients[2,1]*x.rprc
sig.level <- 0.05
qF <- (2/n)*qf(1-sig.level/2, 2, n-2)
meanxy.mean <- mean(meanxy,na.rm=TRUE)
meanxy.sd <- sd(meanxy,na.rm=TRUE)
z2 <- ((x.rprc - meanxy.mean)/meanxy.sd)^2
var.res <- rprc$res_var
y.lwr <- y.rprc - sqrt(qF*(1+z2)*var.res)
y.upr <- y.rprc + sqrt(qF*(1+z2)*var.res)
y.rac0 <- rprc$coefficients[1,1]
z2.0 <- ((-meanxy.mean)/meanxy.sd)^2
y.lwr.0 <- y.rac0 - sqrt(qF*(1+z2.0)*var.res)
y.upr.0 <- y.rac0 + sqrt(qF*(1+z2.0)*var.res)
minx <- min(c(0,x.rprc),na.rm=TRUE)
maxx <- max(c(0,x.rprc),na.rm=TRUE)
```

```

miny <- min(c(0,y.rprc,y.lwr,y.upr),na.rm=TRUE)
maxy <- max(c(0,y.rprc,y.lwr,y.upr),na.rm=TRUE)
sunflowerplot(meanxy,diffyx,
               xlim=c(minx,maxx),ylim=c(miny,maxy),
               main="Test of same precision",
               xlab="(x + y)/2",
               ylab="y - x",
               rotate=TRUE,
               size=.1,
               col="gray",
               seg.col="gray",
               seg.lwd=.8,
               axes=FALSE)
axis(1)
axis(2)
lines(x.rprc, y.rprc, col="darkgray")
lines(x.rprc, y.upr, lty=2)
lines(x.rprc, y.lwr, lty=2)
# H0
abline(h=mean(y,na.rm=TRUE)-mean(x,na.rm=TRUE), lty=3, lwd=0.9)

```

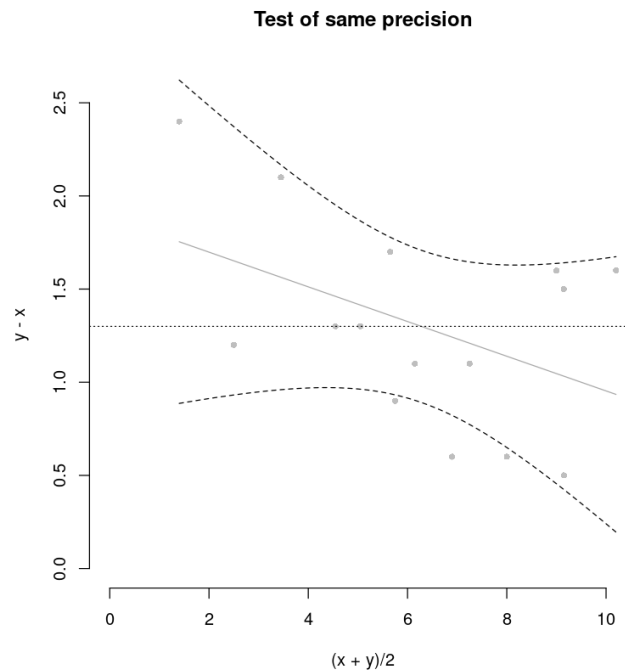


Figure 4: Illustration of the test of same precision. A hypothetical horizontal line can be positioned within the 95% confidence band. In this example, the line was plotted at the level $\bar{y} - \bar{x} = 1.3$.

This inferential approach can be contrasted with the traditional Bland-Altman plot method (Figure 5), which depends on subjective interpretation for equivalence decisions.

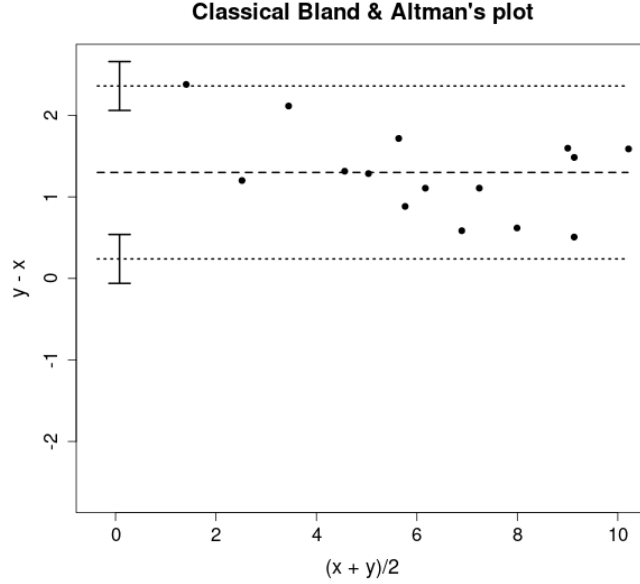


Figure 5: Original Bland-Altman plot showing the traditional interval and the evident bias with $y - x > 0$. This is a Chebyshev interval, which always contains at least 75% of data (compare with Figure 4).

4.5 Test of reliability¹ (Deming regression)

The aim is to test the measurement reliability, which corresponds to the structural null hypothesis that the regression line coincides with the bisector of the first quadrant, i.e., $Y = X$. The main assumption of this structural bisector test for assessing the equivalence of two methods with interval measurements is the absence of bias, as $Y = X$ requires $E[Y] = E[X]$. This test can be performed using Deming regression, which differs from ordinary least squares (OLS) regression in two aspects: it is specifically designed for the bisector line (with the null hypothesis assuming a slope of 1, whereas OLS tests for a slope of 0), and it takes into account measurement errors in both the x and y variables (while OLS assumes that only the y variable has measurement errors).

The null hypothesis is:

$$\begin{aligned} \text{structural } H_0^3 : Y - X &= \alpha + (\beta - 1)X, \alpha = 0, \beta = 1 \\ \text{functional } H_0^3 : y - x &= \alpha + (\beta - 1)x + (\delta - \beta\epsilon), \alpha = 0, \beta = 1 \end{aligned} \quad (16)$$

There is an endogeneity problem:

$$C[x, \delta - \beta\epsilon] = C[X + \epsilon, \delta - \beta\epsilon] = -\beta\sigma_\epsilon^2 < 0 \quad (17)$$

Observe that the explanatory variable x is correlated with the error term $\delta - \beta\epsilon$. Since both x and $\delta - \beta\epsilon$ depend on ϵ , they are correlated, so the OLS estimation of β will be biased downward. Measurement error in the dependent variable y , however, does not cause endogeneity (though it does increase the variance of the error term)¹⁴. Therefore, OLS regression can be performed with the observed variables only if $C[x, \delta - \beta\epsilon] = 0$, if $\beta = 0$ or $\sigma_\epsilon^2 = 0$ (which is another way for the assumption of absence of measurement errors correlated with x).

To estimate the parameters α and β of the structural regression, as well as the structural values X and Y for each observational unit, it is necessary to minimize the following quantity Q ? :

$$Q = \sum_{i=1}^n \left(\frac{\delta_i^2}{\sigma_\delta^2} - \frac{\epsilon_i^2}{\sigma_\epsilon^2} \right) = \frac{1}{\sigma_\delta^2} \sum_{i=1}^n \left((y_i - \alpha - \beta X_i)^2 + \lambda (x_i - X_i)^2 \right) \quad (18)$$

Therefore, assuming there were two observations by each method for each observational unit, i.e., $k = 2$, we have (being a , b , and l the estimators of α , β , and λ , respectively):

$$b = \frac{s_y^2 - ls_x^2 + \sqrt{(s_y^2 - ls_x^2)^2 + 4ls_{xy}^2}}{2s_{xy}} \quad (19)$$

$$a = \bar{y} - b\bar{x}$$

$$l = \frac{s_{yk}^2}{s_{xk}^2}$$

$$s_{xk}^2 = \frac{\sum_{i=1}^n (x_{1i} - x_{2i})^2}{2n}$$

$$s_{yk}^2 = \frac{\sum_{i=1}^n (y_{1i} - y_{2i})^2}{2n}$$

$$x_i = \frac{x_{1i} + x_{2i}}{2}$$

$$y_i = \frac{y_{1i} + y_{2i}}{2}$$

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

$$s_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

$$s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$$

$$s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

The formulas for the estimator of β by McCartin (2003)¹⁵, Isaac (1970)¹⁶, and Wikipedia⁷ are different from Linnet (1998)⁹, Linnet (1990)¹⁷, which is also mentioned in NCSS³, and mcr::mcreg¹⁸.

These authors also adopt the different conventions for the estimation of λ :

- y : Reference method
- x : New method
- $l = \frac{s_{yk}^2}{s_{xk}^2}$: Isaac (1970), McCartin (2003), Wikipedia
- $l = \frac{s_{xk}^2}{s_{yk}^2}$: Shukla (1973), Linnet (1998), NCSS, mcr::mcreg

The estimates of the latent values for both measurements for each observational unit are:

$$\begin{aligned}\hat{X}_i &= x_i + \frac{b}{b^2 + l} (y_i - a - bx_i) \\ \hat{Y}_i &= y_i - \frac{l}{b^2 + l} (y_i - a - bx_i)\end{aligned}\quad (20)$$

The latent and manifest centroids are equal:

$$\begin{aligned}\bar{\hat{X}} &= \frac{\sum_{i=1}^n \hat{X}_i}{n} = \bar{x} \\ \bar{\hat{Y}} &= \frac{\sum_{i=1}^n \hat{Y}_i}{n} = \bar{y}\end{aligned}\quad (21)$$

The estimates of the latent and manifest variances are approximately equal for both measurements:

$$\begin{aligned}s_{\hat{X}}^2 &= \frac{\sum_{i=1}^n (\hat{X}_i - \bar{\hat{X}})^2}{n} \approx s_x^2 \\ s_{\hat{Y}}^2 &= \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}})^2}{n} \approx s_y^2\end{aligned}\quad (22)$$

To test the null hypothesis $\lambda = 1$, the 95% confidence interval of Shukla⁶ is used (see Test of precision (equality of variances), page 12).

Please note that, given the information provided, it is not possible to implement a functional form of OLS regression. Instead, each estimate needs to be calculated using the equations defined earlier. The following is a simplified analytical version of the procedure used in our *eirasagree* package.

To provide a clearer understanding, we will resume the numerical example from Table 1. For this method, the estimation of the lambda parameter is

central (see equation 11). According to Shukla, 1973⁶, it is implemented accordingly to equations 7, 8, and 9, by:

```
data <- na.omit(data.frame(x,y)) # complete data only
n <- nrow(data)
x <- data$x
y <- data$y
vepsilon <- var(x) - cov(x,y, use = "complete.obs")
vdelta <- var(y) - cov(x,y, use = "complete.obs")
gl <- n - 2
sig.level <- 0.05
t <- qt(1 - sig.level/2, gl)
P <- (t^2)*(var(x)*var(y) - cov(x,y, use = "complete.obs")^2)/(n - 2)
L.aux <- rep(NA,2)
L.aux[1] <- abs((vdelta - sqrt(P))/(vepsilon + sqrt(P)))
L.aux[2] <- abs((vdelta + sqrt(P))/(vepsilon - sqrt(P)))
L.aux <- sort(L.aux)
cat("lambda = ",lambda," , IC 97.5% (lambda) = [",1/Lim.lambda[2], " , ",
1/Lim.lambda[1],"]\n",sep="")
```

obtaining:

```
lambda = 6.394012, IC 97.5% (lambda) = [0.0834779, 1.236391]
```

The computation of a (estimate of α) and b (estimate of β), from equation 19, is implemented as:

```
xg <- mean(x, na.rm=TRUE)
yg <- mean(y, na.rm=TRUE)
gl <- n - 2
sig.level <- 0.05
t <- qt(1 - sig.level/2, gl)
sxx <- var(x,na.rm=TRUE)
syy <- var(y,na.rm=TRUE)
sxy <- cov(y,x, use = "complete.obs")
if(sxy=0) {b <- 0} else
{
  b <- (syy - lambda*sxx + sqrt((syy - lambda*sxx)^2 +
4*lambda*(sxy^2)))/(2*sxy)
}
a <- yg - b*xg
r <- cor(x,y,use="complete.obs")
b.se <- b*sqrt((1-r^2)/((r^2)*gl))
lwr.b <- b - t*b.se
upr.b <- b + t*b.se
lwr.a <- a - t*b.se*sqrt(mean(y^2))
upr.a <- a + t*b.se*sqrt(mean(y^2))
cat("alpha = ",a," , IC97.5%(alpha) = [",lwr.a, " , ",upr.a,"]\n",sep="")
cat("beta = ",b," , IC97.5%(beta) = [",lwr.b, " , ",upr.b,"]\n",sep="")
```

resulting in:

```
alpha = 1.86482, IC97.5%(alpha) = [1.121874, 2.607767]
beta = 0.8996172, IC97.5%(beta) = [0.798309, 1.000925]
```

Deming regression also requires, instead of observed values, the estimated true values \hat{X}_i and \hat{Y}_i (Mandel (1984, p. 7, Appendix D)¹⁹) (Table 2), which

are computed by equation 20, implemented by:

```
# estimates of  $\hat{X}$  and  $\hat{Y}$ 
d <- y - (a + b*x)
hatX <- x + (b/(lambda + b^2))*d
hatY <- y - (lambda/(lambda + b^2))*d
```

Table 2: Estimated true values of two measurement methods (x and y) applied to 15 individuals.

individual	\hat{X}	\hat{Y}
1	5.246	6.584
2	8.841	9.818
3	4.385	5.809
4	9.485	10.397
5	1.841	3.521
6	6.525	7.735
7	2.459	4.077
8	3.878	5.354
9	7.639	8.737
10	4.840	6.219
11	6.688	7.882
12	0.269	2.107
13	8.460	9.475
14	5.575	6.880
15	8.270	9.304

The difference between observed and estimated true values is shown in Figure 6.

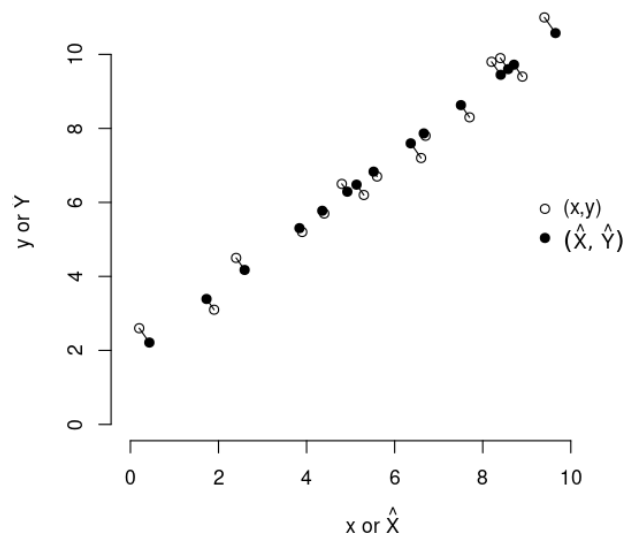


Figure 6: Observed values (x, y) and corresponding true values (\hat{X}, \hat{Y}) estimated by equation 20 (compare with observed values in Table 1).

Finally, with the true values \hat{X} and \hat{Y} , Deming regression can be computed with:

```
# Deming regression
reg.Deming <- estimatr::lm_robust(hatY~hatX)
print(summary(reg.Deming))
```

obtaining:

```
Call:
estimatr::lm_robust(formula = hatY ~ hatX)

Standard error type: HC2

Coefficients:
              Estimate Std. Error  t value  Pr(>|t|)  CI Lower CI Upper DF
(Intercept)  1.8229    9.950e-16  1.832e+15 1.441e-192  1.8229    1.8229  13
hatX         0.9071    2.056e-16  4.411e+15 1.578e-197  0.9071    0.9071  13

Multiple R-squared:  1 ,      Adjusted R-squared:  1
F-statistic: 1.946e+31 on 1 and 13 DF,  p-value: < 2.2e-16
```

Note that we used the function `estimatr::lm_robust` to find the intercept and slope values, but as a statistical test, this regression is useless (in fact, it checks that the computation of a and b , implemented at page 19 and based on equation 19, are correct). The p-values are very small and the confidence

interval limits replicate the point estimate of the intercept and slope. The reason for this is how the Deming regression is calculated. Note that the solid circles in Figure 6, representing the pairs (\bar{X}, \bar{Y}) , are perfectly aligned, resulting in $R^2 = 1$. This perfect alignment means that any imprecision in the numerical rounding becomes statistically significant, leading to the rejection of the null hypothesis $H_0^3 : \alpha = 0, \beta = 1$. The conclusion is that inferential decisions cannot be made based on this analytical procedure.

Another way to express the null hypothesis of equation 16, emphasizing the jointly test of α and β is:

$$\text{structural } H_0^3 : [\alpha \ \beta] = [0 \ 1] \quad (23)$$

One proposal is to calculate a confidence band for the Deming regression, allowing for some statistical uncertainty. As such, to simultaneously test the null hypothesis of intercept and slope in Deming Regression, the Working-Hotelling 95% structural confidence band or the 95% elliptical confidence region for intercept and slope can be employed. However, a challenge with these confidence band and confidence region tests is that the decision regarding the null hypothesis relies on visual inspection. In the case of the structural confidence band, the null hypothesis is not rejected if the bisector line is completely contained within the band limits. As for the structural elliptical confidence region, the point $(0, 1)$, representing the joint null hypothesis $H_0^3 : [\alpha \ \beta] = [0 \ 1]$, must fall within this region.

The 95% structural confidence band of Working-Hotelling (Strike, 1991, chapter 7)²⁰, Kutner et al. (2004, p. 63)²¹, Johnson & Wichern (2007, p. 212)²² is:

$$CB^{95\%}(E[Y|X]) = \left[a + bX \pm Ws_d \sqrt{\frac{1}{n} \left(1 + \left(\frac{X - \hat{X}}{s_{\hat{X}}} \right)^2 \right)} \right]$$

$$X \in [\min\{\hat{X}_i\}, \max\{\hat{X}_i\}]$$

$$W^2 = T^2(2, n - 1) = 2 \frac{n - 1}{n - 2} F_{2, n - 2}^{95\%}$$

$$s_d^2 = \frac{1}{n - 2} \sum_{i=1}^n d_i^2$$

$$d_i = y_i - a - bx_i \quad (24)$$

This confidence band can be implemented by:

```
sig.level <- 0.05
range.x <- seq(min(hatX), max(hatX), length.out=1e3)
a <- reg.Deming$coefficients[1] # or compute a from equation 19
b <- reg.Deming$coefficients[2] # or compute b from equation 19
d <- y - a - b*x
var.res <- (1/(n-2))*sum(d^2)
```

```

qF <- qf(1-sig.level/2, 2, n-2)
W <- sqrt(2*((n-1)/(n-2))*qF)
X.mean <- mean(hatX,na.rm=TRUE)
X.sd <- sd(hatX,na.rm=TRUE)
sq <- W*sqrt(var.res)*sqrt((1/n)*(1+((range.x-X.mean)/X.sd)^2))
range.y <- a + b*range.x
Y.lwr <- range.y - sq
Y.upr <- range.y + sq
min.xy <- min(c(0,hatX,hatY,Y.lwr,Y.upr),na.rm=TRUE)
max.xy <- max(c(0,hatX,hatY,Y.lwr,Y.upr),na.rm=TRUE)
sunflowerplot(hatX,hatY,
               xlim=c(min.xy,max.xy),ylim=c(min.xy,max.xy),
               main="Test of reliability (Deming regression)",
               xlab="X = True[x]",
               ylab="Y = True[y]",
               rotate=TRUE,
               size=.1,
               col="gray",
               seg.col="gray",
               seg.lwd=.8,
               axes=FALSE)
axis(1)
axis(2)
curve(a + b*x, from=min(hatX), to=max(hatX),
      col="darkgray", add=TRUE)
lines(range.x, Y.upr, lty=2)
lines(range.x, Y.lwr, lty=2)
curve(x^1, from=0, to=max.xy, lty=3, lwd=0.9, add=TRUE)

```

The resulting graphical representation is shown in Figure 7. The implementation of this analytical confidence band is practically the same as the bands used for the accuracy and same precision tests, with the exception of the variable `var.res` (residual variance). In the previous scripts, `var.res` utilized the calculation from `estimatr:lm_robust`, but here it is replaced by $\frac{\sum d^2}{n-2}$, where d is obtained from equation 20 (implemented on page 20).

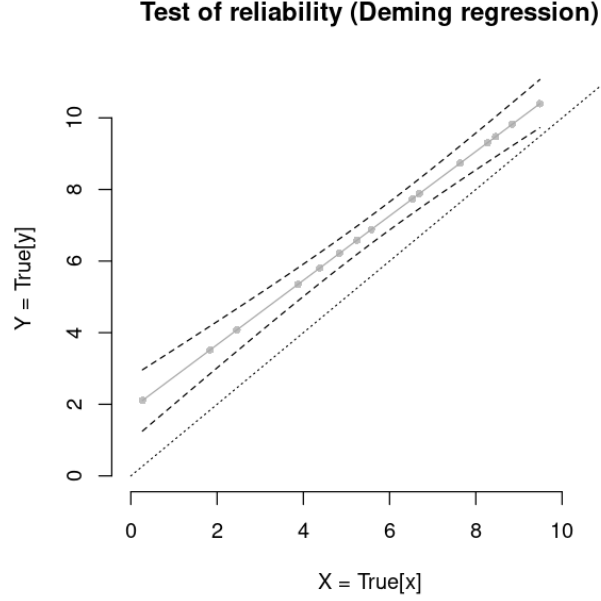


Figure 7: Illustration of Deming regression with a 95% confidence band. The estimated true values (\hat{X}, \hat{Y}) are perfectly aligned and represented as gray circles. The dashed black lines represent the limits of the confidence band, and the dotted black line is the bisector. In this example, the confidence band does not include the bisector, indicating a rejection of the null hypothesis of reliability between measurement techniques. However, it can be observed that the regression line is nearly parallel to the bisector, but with a displacement due to the bias.

The 95% structural elliptical confidence region for intercept and slope (Pestana e Gageiro (2005, p. 76-77)²³) is:

$$\frac{n-2}{2} \frac{n(\alpha-a)^2 + 2(\alpha-a)(\beta-b) \sum_{i=1}^n \hat{X}_i + (\beta^2-b) \sum_{i=1}^n \hat{X}_i^2}{\sum_{i=1}^n d_i^2} \leq F_{2,n-2}^{95\%}$$

$$d_i = y_i - a - bx_i \quad (25)$$

As an ellipse is not a function but a relation, we can represent it in the form of two functions that trace two halves, computing the slope as a function of the

intercept or vice versa. The pair of functions is given by:

$$\beta = b + \frac{\frac{b}{n} + (\alpha - a)\bar{\hat{X}} \pm \sqrt{\frac{2}{n} F_{2,n-2}^{95\%} s_d^2 \bar{\hat{X}}^2 - (\alpha - a)^2 s_{\hat{X}}^2}}{\bar{\hat{X}}^2}$$

$$s_{\hat{X}}^2 = \frac{1}{n} \sum_{i=1}^n (\hat{X}_i - \bar{\hat{X}})^2$$

$$s_d^2 = \frac{1}{n-2} \sum_{i=1}^n d_i^2$$

$$d_i = y_i - a - bx_i \quad (26)$$

The graphical representation of the equation 26 for a 95% ellipse confidence region is depicted in Figure 8 and can be implemented as:

```
sig.level <- 0.05
range.a <- seq(min(c(0,y,hay),na.rm=TRUE),
              max(c(0,y,hay),na.rm=TRUE),
              length.out=1e4)
mean.X <- mean(hatX,na.rm=TRUE)
mean.X2 <- mean(hatX^2,na.rm=TRUE)
d <- y - a - b*x
s2 <- sum(d^2,na.rm=TRUE)/(n-2)
qF <- qf(1-sig.level/2, 2, n-2)
curr.w <- getOption("warn") # avoid warning with sqrt(<0)
options(warn = -1)
sq <- sqrt( ((a - range.a)^2) * ((mean.X^2)-mean.X2) +
           (2/n)*qF*s2*mean.X2 )
options(warn = curr.w) # restore current warning level
numerator1 <- b/n - (range.a - a) * mean.X - sq
numerator2 <- b/n - (range.a - a) * mean.X + sq
dt_tmp1 <- data.frame(range.a,numerator1,rep(1,length(range.a)))
dt_tmp2 <- data.frame(range.a,numerator2,rep(2,length(range.a)))
names(dt_tmp1) <- names(dt_tmp2) <- c("intercept","numerator","part")
dt_ellipse <- rbind(dt_tmp1,dt_tmp2)
dt_ellipse <- dt_ellipse[is.finite(dt_ellipse$numerator),]
rm(numerator1)
rm(numerator2)
rm(dt_tmp1)
rm(dt_tmp2)
dt_ellipse$slope <- b+dt_ellipse$numerator/mean.X2
r1.aux <- which(dt_ellipse$part==1)
r2.aux <- which(dt_ellipse$part==2)
dt_ellipse <- dt_ellipse[c(r1.aux,rev(r2.aux)),]
dt_ellipse <- rbind(dt_ellipse,dt_ellipse[1,])
dt_ellipse$numerator <- NULL
dt_ellipse$part <- NULL
min.x <- min(c(1,dt_ellipse$slope),na.rm=TRUE)-0.1
max.x <- max(c(1,dt_ellipse$slope),na.rm=TRUE)+0.1
min.y <- min(c(0,dt_ellipse$intercept),na.rm=TRUE)-0.1
max.y <- max(c(0,dt_ellipse$intercept),na.rm=TRUE)+0.1
plot(dt_ellipse$slope,
     dt_ellipse$intercept,
     type="l", axes=FALSE,
     xlab="slope", xlim=c(min.x,max.x),
     ylab="intercept", ylim=c(min.y,max.y))
axis(1)
axis(2)
points(1,0,pch=19)
text(1,0,"H0",pos=4)
```

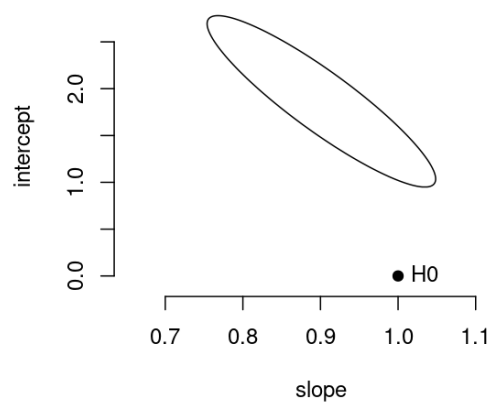


Figure 8: Illustration of 95% confidence ellipse, which does not include $H_0^3 : [\alpha \ \beta] = [0 \ 1]$ (thus, coherently with Figure 7, it is also rejecting the coincidence with bisector).

5 Translations

As can be observed in the numerical example that illustrated the demonstrations of the theorems providing statistical support for the equivalence of accuracy, precision, and reliability between individual measurements, the accuracy test shows a statistically significant bias, with method y overestimating the measures compared to x (Figure 1). Despite this bias, the precision test (Figure 3) does not reject the null hypothesis because this test only evaluates the slope of the regression and it is possible to accommodate a horizontal line within the limits of the 95% confidence bands. However, considering $y - x$ as the dependent variable, if there were no bias, this horizontal line should be at the null value ($y - x = 0$), which is entirely outside the confidence bands. Finally, it is worth noting that the Deming regression test rejected the null hypothesis because the bisector line is outside the band (Figure 7) and because the ordered pair (0,1) is outside the ellipse (Figure 8). This is a counterintuitive rejection, and it can be observed that a line with a slope of 1, parallel to the bisector, could be accommodated within the confidence band.

Considering that bias is likely the most straightforward issue to address when replacing a measurement technique, it would be imprudent not to investigate whether a simple addition or subtraction to correct this bias could salvage the candidate technique. To facilitate this analysis, the `eirasagree` package has been enhanced by default to include the computation of these tests with automatic bias correction. To ensure robustness, the package also offers the corresponding 95% confidence band and elliptical region using bootstrapping techniques instead of the analytical approach described above. These corrections provide additional insights and support a more comprehensive assessment of the measurement techniques.

The descriptive statistics and regression results, could be obtained simply using the following command:

```
x <- c(5.3, 8.9, 4.4, 9.4, 1.9, 6.6, 2.4, 3.9, 7.7, 4.8, 6.7, 0.2, 8.4, 5.6, 8.2)
y <- c(6.2, 9.4, 5.7, 11.0, 3.1, 7.2, 4.5, 5.2, 8.3, 6.5, 7.8, 2.6, 9.9, 6.7, 9.8)
data <- data.frame(x,y)
output <- eirasagree::AllStructuralTests(data,
                                         reference.cols=1,
                                         newmethod.cols=2,
                                         alpha=0.05, out.format="html",
                                         B = 5e3)
```

The resulting tests shown in Figure 9 employ a bootstrap approach, where decisions are made based on the positioning of elements within the boundaries of bands or ellipses. In this example, the test for accuracy in Figure 9 (top panel) resulted in rejection, indicating a deviation from equal accuracy. However, when considering the 95% confidence interval of the bias, which is approximately [1.02, 1.54], the test for precision was conducted (Figure 9, middle panel). It shows that the line representing the null hypothesis of $y - x = 0$ lies outside the 95% confidence band. Nevertheless, attempts to fit horizontal lines within the range of the bias interval were successful in not rejecting the equivalence of precision.

Similarly, in the reliability test (Figure 9, bottom panel), the null hypothesis is presented on the bisector, and the parallel lines are also adjusted according to the bias interval range. These adjustments led to successful outcomes. The concurrent test using the confidence ellipse, which estimates the intercept displaced by the bias range, also yielded successful results. Using translation may resolve the issue of counterintuitive rejections mentioned above.

Alternatively, by setting $B=0$, analytical methods are applied instead of bootstrap methods, which is depicted in Figure 10.

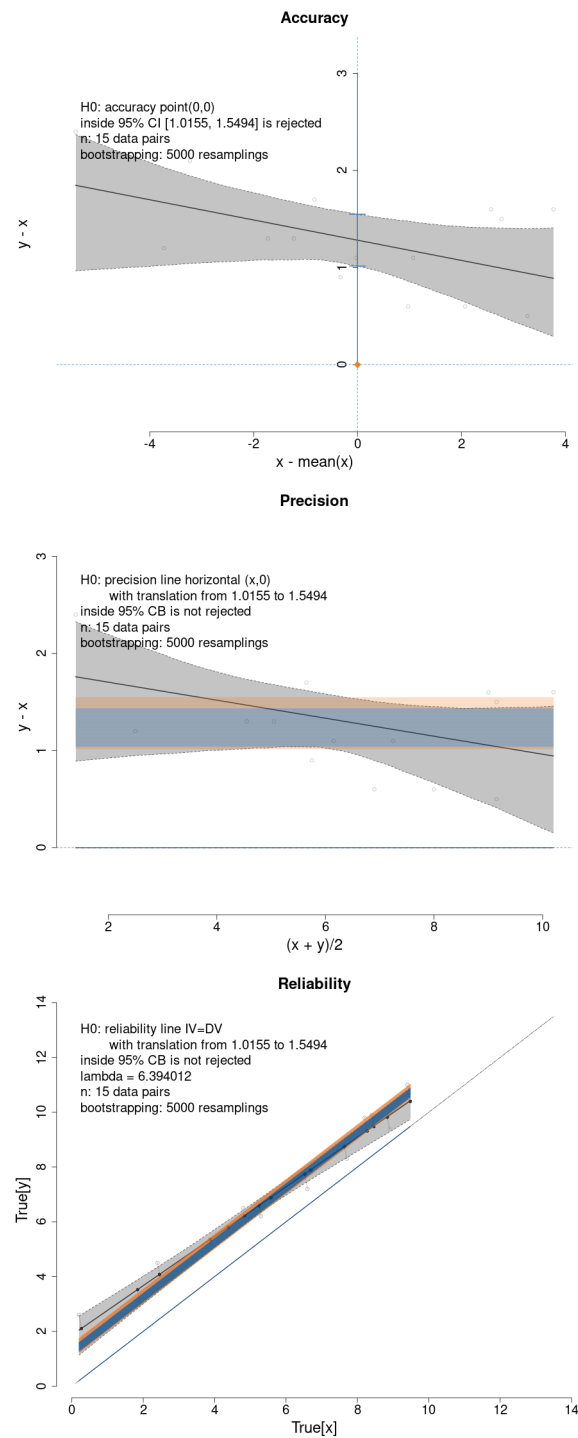


Figure 9: Graphical output from package eirasagree obtained with 5×10^3 bootstrap iterations. Top panel: accuracy (compare with Figures 1 and 2). Middle panel: precision (compare with Figures 3 and 4). Bottom panel: agreement with bisector (compare with Figures 7 and 8).

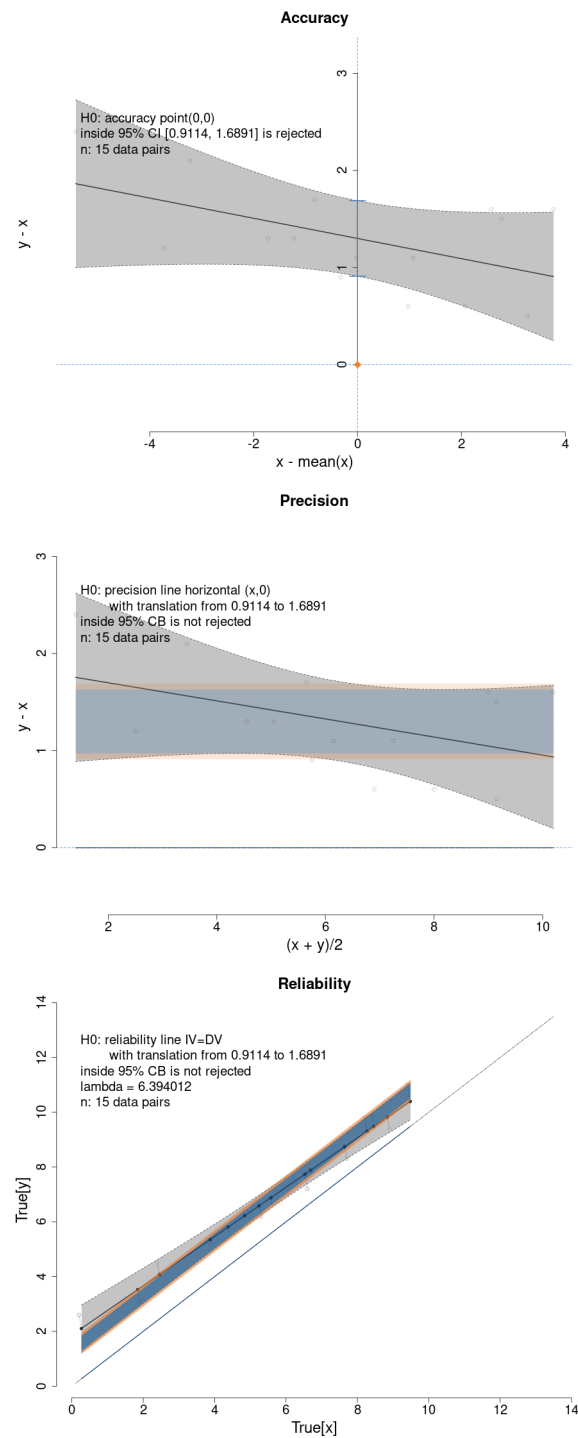


Figure 10: Graphical output from package eirasagree, in the analytic alternative to Figure 9. Top panel: accuracy (compare with Figures 1 and 2). Middle panel: precision (compare with Figure 3 and 4). Bottom panel: agreement with bisector (compare with Figures 7 and 8).

6 Getting started with eirasagree

In order to generate all the statistics and graphs shown in Figure 9, our package simplifies the necessary R work to a great extent. It only required:

```
data <- eirasdata::Demonstration
eirasagree::AllStructuralTests(data,
                                reference.cols=1,
                                newmethod.cols=2,
                                alpha=0.05, out.format="txt",
                                B = 5e3)
```

In this integrative function, **data** can receive any data frame with one or more columns for each measure. For instance, **data** can be replaced by:

```
data <- readxl::read_excel("mydata.xlsx")
```

in which, `mydata.xlsx` is an Excel worksheet with two numeric columns.

This also work:

```
x <- c(5.3, 8.9, 4.4, 9.4, 1.9, 6.6, 2.4, 3.9, 7.7, 4.8, 6.7, 0.2, 8.4, 5.6, 8.2)
y <- c(6.2, 9.4, 5.7, 11.0, 3.1, 7.2, 4.5, 5.2, 8.3, 6.5, 7.8, 2.6, 9.9, 6.7, 9.8)
data <- data.frame(x,y)
```

By setting the parameter `out.format="txt"`, the package saves on file what it displays on the screen while keeping the images available in a separate folder. With `out.format` set to `"html"` or `"pdf"`, the package automatically generates a report that integrates both the text and figures into a single document. In all formats, figures are stored in `analysis/graph` (from which Figure 9 was composed). A typical textual output with `out.format="txt"` and `B=5e3` reports:

```
-----
x and y
-----

-----
Descriptive statistics
-----

-----
- data under analysis
-----

      x    y
1  5.3  6.2
2  8.9  9.4
3  4.4  5.7
4  9.4 11.0
5  1.9  3.1
6  6.6  7.2
7  2.4  4.5
8  3.9  5.2
9  7.7  8.3
10 4.8  6.5
11 6.7  7.8
12 0.2  2.6
13 8.4  9.9
14 5.6  6.7
```

```

15  8.2  9.8

-----
- summary
-----

```

	x	y
Min.	Min.0.200	Min.2.600
1st Qu.	1stQu.4.150	1stQu.5.450
Median	Median5.600	Median6.700
3rd Qu.	3rdQu.7.950	3rdQu.8.850
Max.	Max.9.400	Max.11.000
Mean	Mean5.627	Mean6.927
s.d.	2.735	2.493
n	15	15
NA	0	0

```

-----
Traditional correlation
-----

Pearson's product-moment correlation

data: as.numeric(unlist(data[, 1])) and as.numeric(unlist(data[, 2]))
t = 19.184, df = 13, p-value = 6.42e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9475952 0.9944178
sample estimates:
      cor
0.9827929

-----
Classical Bland-Altman estimates
-----

```

	estimate
mean.diff	1.300000
lowerLoA	0.239321
upperLoA	2.360679
ciLoA	0.299686
alpha	0.050000

```

-----
Test of structural accuracy
-----
Hedberg, EC, Ayers, S (2015) The power of a paired t-test
with a covariate. Social Science Research 50: 277-91

-----
- model
-----
      IV = x - mean(x)
      DV = y - x

-----
- functional robust approach
-----

Decision by 95% confidence interval:
      avg{y - x} = 1.2821 + -0.1041 {x - mean(x)}

      H0: accuracy point(0,0) inside 95% CI [1.0155, 1.5494] is rejected
      n: 15 data pairs
      bootstrapping: 5000 resamplings
      Bias 95%CI: between 1.0155 and 1.5494.

-----
Test of structural precision
-----

```

```

-----
- lambda test
-----

Shukla, GK (1973) Some exact tests on hypothesis
about Grubbs estimators. Biometrics 29: 373-377
https://doi.org/10.2307/2529399

Assuming:
- reference method: x
- putative method: y

lambda = V[delta]/V[epsilon] = 1 is not rejected

Shukla, GK (1973) Some exact tests on hypothesis
about Grubbs estimators. Biometrics 29: 373-377

-----
- model
-----
      IV = (x + y) / 2
      DV = y - x

-----
- functional robust approach
-----

Decision by 95% confidence band:
      avg{y - x} = 1.89 + -0.0928 {(x + y)/2}

      H0: precision line horizontal (x,0)
      with translation from 1.0155 to 1.5494 inside 95% CB is not rejected
      n: 15 data pairs
      bootstrapping: 5000 resamplings

-----
structural bisector line
-----

Creasy, MA (1956) Confidence Limits for the Gradient
in the Linear Functional Relationship. Journal of the
Royal Statistical Society 18(1):65-69

Glaister, P (2001) Least squares revisited.
The Mathematical Gazette 85(502): 104-107.

-----
- model
-----
      X = True[x]
      Y = True[y]

Deming regression for reliability
      Y = intercept + slope . X

-----
- functional robust approach
-----

Decision by 95% confidence band:
      avg{True[y]} = 1.865 + 0.902 {True[x]}

      H0: reliability line IV=DV
      with translation from 1.0155 to 1.5494 inside 95% CB is not rejected
lambda = 6.394012
      n: 15 data pairs
      bootstrapping: 5000 resamplings
-----

```

```

structural confidence elliptical region
-----

Decision by 95% confidence ellipse:

      H0: intercept = [1.0155,1.5494], slope = 1 inside 95% ellipse is not rejected
      n: 15 data pairs
      bootstrapping: 5000 resamplings

-----

See also figures (png format) available in subfolder "graph"
-----

```

For comparison, with `out.format="txt"` and `B=0` the textual output would report:

```

-----
x and y
-----

-----
Descriptive statistics
-----

-----
- data under analysis
-----

      x      y
1    5.3    6.2
2    8.9    9.4
3    4.4    5.7
4    9.4   11.0
5    1.9    3.1
6    6.6    7.2
7    2.4    4.5
8    3.9    5.2
9    7.7    8.3
10   4.8    6.5
11   6.7    7.8
12   0.2    2.6
13   8.4    9.9
14   5.6    6.7
15   8.2    9.8

-----
- summary
-----

      x      y
Min.   Min.0.200  Min.2.600
1st Qu. 1stQu.4.150 1stQu.5.450
Median Median5.600  Median6.700
3rd Qu. 3rdQu.7.950 3rdQu.8.850
Max.   Max.9.400  Max.11.000
Mean   Mean5.627  Mean6.927
s.d.   2.735    2.493
n      15      15
NA     0      0

-----
Traditional correlation
-----

Pearson's product-moment correlation

```

```

data: as.numeric(unlist(data[, 1])) and as.numeric(unlist(data[, 2]))
t = 19.184, df = 13, p-value = 6.42e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9475952 0.9944178
sample estimates:
      cor
0.9827929

-----
Classical Bland-Altman estimates
-----
      estimate
mean.diff 1.300000
lowerLoA  0.239321
upperLoA  2.360679
ciLoA     0.299686
alpha     0.050000

-----
Test of structural accuracy
-----
Hedberg, EC, Ayers, S (2015) The power of a paired t-test
with a covariate. Social Science Research 50: 277-91

-----
- model
-----
      IV = x - mean(x)
      DV = y - x

-----
- functional analytic approach
-----

Call: estimatr::lm_robust(formula = DV ~ IV)

Standard error type:      HC2

Coefficients:
      Estimate  Std. Error  t value  Pr(>|t|)  CI Lower  CI Upper  DF
(Intercept)    1.300      0.126    10.329   0.000    1.028    1.57213.000
IV             -0.104      0.058    -1.779   0.099   -0.230    0.02213.000

Multiple R-squared: 0.2758, Adjusted R-squared: 0.2201
F-statistic: 3.2 on 1 and 13 DF, p-value: 0.098567

Decision (analytic):

      H0: equal structural means between methods is rejected
      (p = 1.23e-07)

Decision by 95% confidence interval:
      avg{y - x} = 1.3 + -0.1039 {x - mean(x)}

      H0: accuracy point(0,0) inside 95% CI [0.9114, 1.6891] is rejected
      n: 15 data pairs
      Bias 95%CI: between 0.9114 and 1.6891.

-----
Test of structural precision
-----

-----
- lambda test
-----

```

```

Shukla, GK (1973) Some exact tests on hypothesis
about Grubbs estimators. Biometrics 29: 373-377
https://doi.org/10.2307/2529399

Assuming:
  - reference method: x
  - putative method: y

lambda = V[delta]/V[epsilon] = 1 is not rejected

Shukla, GK (1973) Some exact tests on hypothesis
about Grubbs estimators. Biometrics 29: 373-377

-----
- model
-----
      IV = (x + y) / 2
      DV = y - x

-----
- functional analytic approach
-----

Call: estimatr::lm_robust(formula = DV ~ IV)

Standard error type:      HC2

Coefficients:
      Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
(Intercept)  1.884    0.427   4.418   0.001   0.963  2.80613.000
      IV      -0.093    0.066  -1.419   0.179  -0.235  0.04913.000

Multiple R-squared: 0.2005, Adjusted R-squared: 0.139
F-statistic: 2 on 1 and 13 DF, p-value: 0.179407

Decision (analytic):

      H0: equal structural precisions between methods is not rejected
      (p = 0.1794)

Decision by 95% confidence band:
      avg{y - x} = 1.8844 + -0.0931 {(x + y)/2}

      H0: precision line horizontal (x,0)
      with translation from 0.9114 to 1.6891 inside 95% CB is not rejected
      n: 15 data pairs

-----
structural bisector line
-----

Creasy, MA (1956) Confidence Limits for the Gradient
in the Linear Functional Relationship. Journal of the
Royal Statistical Society 18(1):65-69

Glaister, P (2001) Least squares revisited.
The Mathematical Gazette 85(502): 104-107.

-----
- model
-----

      X = True[x]
      Y = True[y]

Deming regression for reliability
      Y = intercept + slope . X

```

```

-----
- functional analytic approach
-----

Decision by 95% confidence band:
  avg{True[y]} = 1.8648 + 0.8996 {True[x]}

  H0: reliability line IV=DV
  with translation from 0.9114 to 1.6891 inside 95% CB is not rejected
lambda = 6.394012
  n: 15 data pairs
-----

structural confidence elliptical region
-----

Decision by 95% confidence ellipse:

  H0: intercept = [0.9114,1.6891], slope = 1 inside 95% ellipse is not rejected
  n: 15 data pairs

-----

See also figures (png format) available in subfolder "graph"
-----

```

The difference with the analytical methods is the inclusion of a p value for inferential decision. However, in both cases, bootstrap or analytical, the decision is based on the inclusion of points or lines within the band limits. For that, the package incorporates an automated visual inspection feature that checks for the presence of points within the respective areas, providing a statistical decision without requiring researcher interpretation. This feature is particularly useful when the null hypothesis is close to the limit of a confidence band or elliptical region. It is also important to note that there is no analytical decision by p value for Deming regression; for this test, the decision is always based on inclusion in the areas.

In addition to the inferential statistics, the graphs depicted in Figure 11 are also available. Other resources and alternatives for the package applications are detailed in the package documentation (see “Instructions for package installation”, page 2 and “A first example”, page 3).

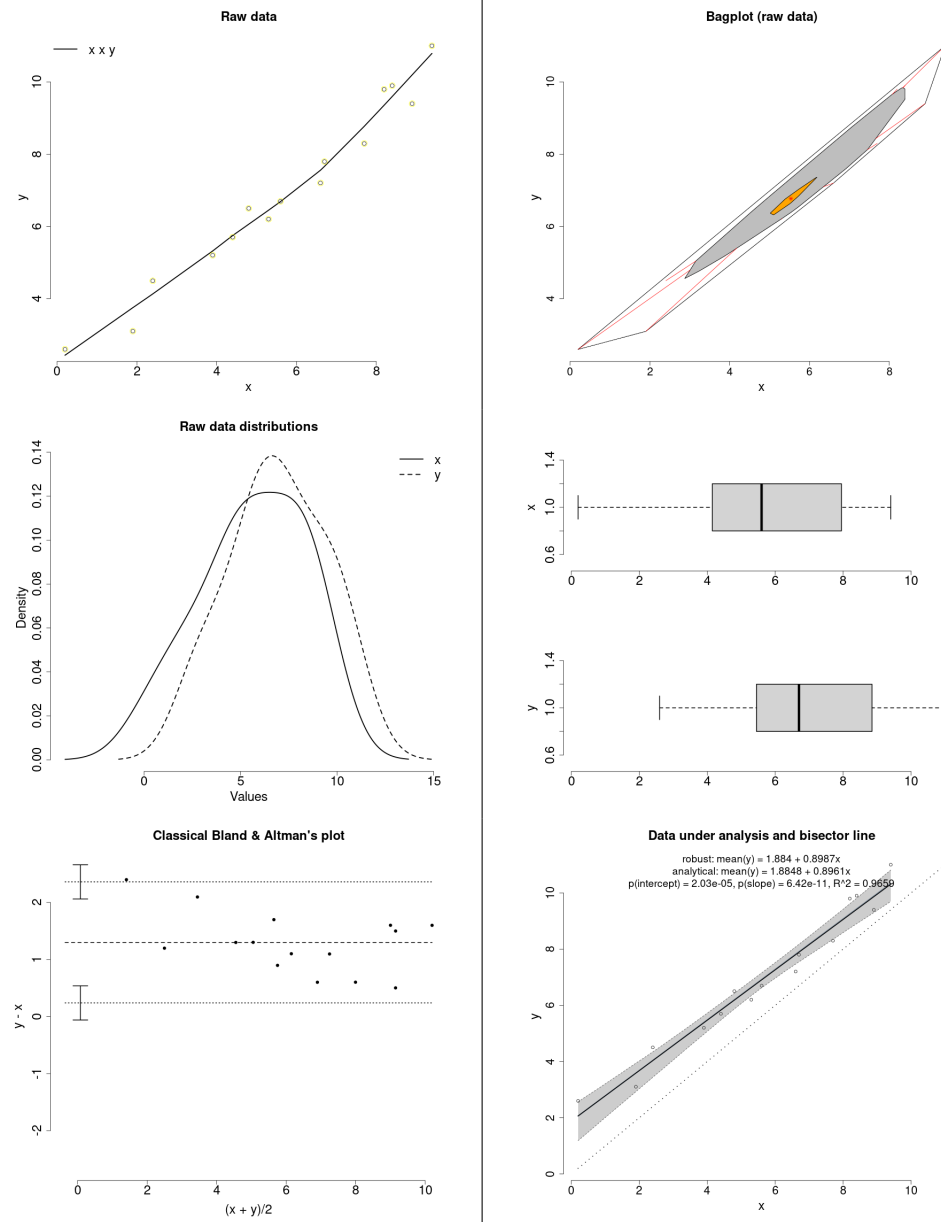


Figure 11: Examples of other graphs available with eirasagrec.

References

1. Traub R, Rowley G. Understanding Reliability. Instructional Topics in Educational Measurement. 1992.
2. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*. 1986 6;327(8476):307-10.
3. NCSS Statistical Software [book on the Internet]. Deming regression;. Available from: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Deming_Regression.pdf.
4. Creasy MA. Confidence Limits for the Gradient in the Linear Functional Relationship. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1956;18(1).
5. Hedberg EC, Ayers S. The power of a paired t-test with a covariate. *Social Science Research*. 2015;50.
6. Shukla GK. Some Exact Tests of Hypotheses about Grubbs's Estimators. *Biometrics*. 1973 6;29(2):373.
7. Glaister P. 85.13 Least squares revisited. *The Mathematical Gazette*. 2001;85(502).
8. Oldham PD. A note on the analysis of repeated measurements of the same subjects. *Journal of Chronic Diseases*. 1962;15(10).
9. Linnet K. Performance of Deming regression analysis in case of misspecified analytical error ratio in method comparison studies. *Clinical Chemistry*. 1998;44(5).
10. Kummell CH. Reduction of Observation Equations Which Contain More Than One Observed Quantity. *The Analyst*. 1879;6(4).
11. Albert A. Statistical methods in laboratory medicine.. vol. 11. Strike PW, editor. Butterworth-Heinemann Ltd., Oxford,; 1991.
12. Watson PF, Petrie A. Method agreement analysis: A review of correct methodology; 2010.
13. Shoukri MM. Measures of Interobserver Agreement and Reliability; 2010.
14. Thoresen M, Laake P. On the simple linear regression model with correlated measurement errors. *Journal of Statistical Planning and Inference*. 2007;137(1).
15. McCartin BJ. A geometric characterization of linear regression. *Statistics*. 2003;37(2).
16. Isaac PD. Linear regression, structural relations, and measurement error. *Psychological Bulletin*. 1970;74(3).

17. Linnet K. Estimation of the linear relationship between the measurements of two methods with proportional errors. *Statistics in Medicine*. 1990;9(12).
18. Potapov S, Model F, Schuetzenmeister A, Manuilova E, Dufey F, Raymaekers J. mcr: Method Comparison Regression; 2023. Available from: <https://CRAN.R-project.org/package=mcr>.
19. Mandel J. FITTING STRAIGHT LINES WHEN BOTH VARIABLES ARE SUBJECT TO ERROR. *Journal of Quality Technology*. 1984;16(1).
20. Strike PW. *Statistical Methods in Laboratory Medicine*; 1991.
21. Kutner M, Nachtsheim C, Neter J, Li W. *Applied Statistical Linear Models*. McGraw-Hill Companies I, editor; 2005.
22. Johnson RA, Wichern DW. *Applied Multivariate Statistical Analysis.*: Pearson Prentice Hall; 2007.
23. Pestana M, Gageiro J. *Descobrimdo a regressão*. Lisboa: Sílabo; 2005.